EXPLAINING BLACK-BOX MACHINE LEARNING MODELS

Krzysztof Dzieciolowski

*John Molson School of Business, Concordia University
Montreal, Canada*

ABSTRACT

Despite the advances of Machine Learning (ML), the models are still being considered black-boxes that are difficult to explain. In order to address the need to interpret the ML models, I suggest the novel Input Prediction Association (IPA) plots which are based on the means of inputs and predictions, calculated in the model's prediction ranks. The resulting plots provide a model-agnostic method of visualizing associations between the inputs and predictions. The IPA plots can help detect inputs that are linearly correlated with predictions as well as help identify the pattern of their relationships. The IPA plots can also be used to establish inputs' thresholds that correspond to high predicted probabilities for the event of interest, thus extending the interpretation of the model inputs to an individual observation level. Such an approach can be applied in Customer Relationship Management since it suggests not only who to target but also it provides insights as to why to target specific individuals. We illustrate the IPA approach with an example from the field of home equity loan risk analysis in banking.

KEYWORDS

Machine Learning, Black-box models, Inputs, Predictions

1. INTRODUCTION

Predicting the probability of customer events is an important task in the application of machine learning (ML) models in the business environment. The ML models need not only to predict accurately and robustly which customers will, for example, likely purchase a product, but also explain why they would do it. In the paper, I suggest an Input Prediction Association (IPA) approach to plot model's ranks inputs means against the means of the predictions. The value of an IPA plots is that it may help discover the pattern of the relationship between the inputs and model predictions. IPA plot may also be used to customize individual recommendations in Customer Relationship Management (CRM) applications.

The discussion about the explanatory and predictive approaches to data-modelling initiated by Breiman (2001) and continued by Shmueli (2010) and others have helped create a focus on creating more accurate and robust predictive models. However, the machine learning models are still the black-boxes that cannot be easily interpreted. There has been growing interest in finding ways to interpret the models, most notably by Friedman (2001) who suggested the heuristic Partial Dependency Plots (PDP) and by Goldstein et al (2015) who proposed using Individual Conditional Expectation Expectations (ICE). Implementation of PDP and ICE plots in SAS software has been developed by Wright (2018). Zhao et al. (2019) have shown that these plots can help extract causal relationships from the black-box models. However, the PDP approach can work only when the variables of interest and non-interest are independent. Both PDP and ICE plots are also a subject to high variability, which may obscure the relationship between the inputs and predictions. In order to address this question, I suggest the Input Prediction Association (IPA) plots which provide the visual, model-agnostic approach of assessing the relationship between inputs and predictions. The IPA approach is different from PDP method in that, instead of integrating-out the inputs of no interest, the inputs of interest are conditioned on the prediction values. In section 2, I define the IPA approach and provide a heuristic algorithm for constructing IPA plots. An illustrative example of IPA plots for Home Equity Loans data is provided there too. Conclusions and recommendations for further research follow in section 3.

2. INTERPRETING BLACK-BOX MODEL WITH IPA PLOTS

In order to visualize the effect of a single input variable on model predictions, I suggest graphing the means of the inputs and predictions obtained from model ranks. The resulting plot is called an Input Prediction Association (IPA) plot. The novelty of the IPA plot is that it is obtained directly from the input variables and model predictions conditioned on the model ranks. The IPA plot helps to identify ranks-based “smoothed-out” relationships between the inputs and predictions. If the goal of the analysis is to extract linear relationships, then Pearson correlations can be calculated from the IPA plots. Here are the steps to obtain an IPA plot.

Step A. Construct a model to predict the target response Y using the inputs $X = (X_1, \dots, X_p)$. Rank-order the model predictions in k quantiles (e.g. $k=20$).

Step B. Calculate the k means for each of p input variables to get $\bar{X}_{ij}$ ($i \leq p, j \leq k$) and for predicted values to get $\bar{Y}_j$ ($j \leq k$). Plot $\bar{X}_{ij}$ versus $\bar{Y}_j$ in k -ranks ($j \leq k$) to obtain an IPA plot for an i^{th} input ($i \leq p$).

Step C. For detecting a linear relationship between an i^{th} input and predictions, calculate the Pearson correlation coefficient, $r_i = \text{Corr}(\bar{X}_i, \bar{Y})$ from k ranks. The inputs that are linearly associated with model predictions can be identified as having high values of $r_i, i \leq p$. In regression models, the slopes of IPA plots are proportional to its leverage of the “hat” projection matrix and corresponding regression coefficient. Let’s assume a linear regression model, $y_m = \sum_{j=1}^p (\beta_0 + \beta_j x_{mj}) + \varepsilon_m$, where $m \leq n$, p is the number of inputs and n is the number of observations. In order to explain the effect of the input x_j on $\hat{y}_m$, using the chain rule, we calculate the partial derivative of $\hat{y}_m$ with respect to input x_j ,

$$\frac{\partial \hat{y}_m}{\partial x_j} = \frac{\partial \hat{y}_m}{\partial y_m} \frac{\partial y_m}{\partial x_j} = h_{mm} \beta_j$$

where h_{mm} is an m^{th} diagonal element of hat projection matrix $H = X(X'X)^{-1}X'$ and $\hat{y} = Hy$. Hence, in the linear regression model, an input x_j “influences” the prediction $\hat{y}_m$ through m^{th} leverage and j^{th} slope.

3. EXAMPLE

We illustrate the construction of IPA plots defined in section 2 using the example of Home Equity Loan data ($n=5960$) from the library SAMPSIO of SAS® Enterprise Miner™ software.

Step A. Obtained a Gradient Boosting model to predict the likelihood of the binary response, “default or seriously delinquent” ($\text{bad}=1$), using 10 inputs, as shown in Table 1.

Table 1. Inputs and IPA correlations for Bank Home Equity loan data

Input number	Input Description	IPA Rank Correlations (sorted)	Population Correlations	Gradient Boosting Importance
1	Number of major derogatory reports	0.98	0.44	1
2	Number of delinquent trade lines	0.95	0.60	10
3	Number of recent credit inquiries	0.95	0.31	5
4	Debt to income ratio	0.79	0.25	2
5	Age of oldest trade line in months	0.66	0.23	8
6	Amount due on existing mortgage	0.53	0.08	4
7	Amount of current loan request	0.47	0.09	3
8	Years on current job	0.35	0.08	7
9	Value of current property	0.30	0.05	6
10	Number of trade (credit) lines	0.05	0.01	9

Step B. Calculate means of inputs and predictions in ($k=20$) ranks and their Pearson correlation coefficients.

Note that, as shown in the Table 1, the input “Number of major derogatory reports” has the highest correlation, $r_1=0.98$, with predictions, followed by the “Number of delinquent trade lines”, $r_2=0.95$. Observe also, that correlations of the variables in the ranks are much higher than those in the population, due to variance reduction resulting from averaging the input and predictions in the ranks. Such noise-canceling approach helps to reveal the significant relationships. In addition, the IPA based inputs’ importance is quite different from that of gradient boosting model. Hence IPA plots may detect important variables, not noticed by the model.

Step C. Construct Input Predictions Association (IPA) plots for Bank Home Equity Loans variables

The IPA plots for the three inputs from Step B are shown in Figure 1 where the means of inputs in the ranks are plotted against the corresponding means of predicted probabilities of loan default ($\text{bad}=1$). The high likelihood of defaulting on home equity loan suggests the high values of “Number of major derogatory reports” and “Number of delinquent trade lines”. However, the “Value of current property” reveals a quadratic-like relationship with predictions. Using a high threshold value for the loan default prediction may help identify customers who satisfy the threshold value of the corresponding inputs. These inputs can be used to provide insights as recommendations for customer-facing agents in CRM applications.

Figure 1. Input Prediction Association (IPA) plots for Home Equity Loans data (default on loan $\text{bad}=1$)

4. CONCLUSIONS

The IPA plot is a visual, model-agnostic tool to help reveal the nature of the relationship between inputs and predictions. In addition, the individual observations can be classified as satisfying input-based threshold criteria to be qualified for the targeted campaign. Hence, IPA plots may help understand black-box ML models both at the population and individual customer levels. It could be useful to find an optimal number of ranks, k , since as $k \rightarrow n$, IPA rank correlations between inputs and predictions converge to population correlations, which are not helpful. Addressing these questions would help explain ML black-box models and advance their applications in CRM applications.

ACKNOWLEDGEMENT

Author would like to thank Concordia University Part-Time Faculty Association (CUPFA) for the financial support of this research.

REFERENCES

- Breiman, L. 2001, Statistical modeling: The two cultures. *Statistical Science*, Vol. 16, 199–215.
- Friedman, J.H. (2001), Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29
- Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2014). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *arXiv:1309.6392v2*.
- Shmueli G., 2010, To Explain or To Predict? *Statistical Science*, Vol. 25, No. 3, 2010, 289–310
- Wright R., 2018. Interpreting Black-Box Machine Learning Models Using PDP and ICE Plots, *SAS Institute Inc.*
- Zhao Q., Hastie T., 2019. Causal Interpretation of Black-Box Models. *Journal of Business and Economic Statistics*